2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Data Gap Modeling in Continuous Glucose Monitoring Sensor Data

Martina Drecogna, Martina Vettoretti, Simone Del Favero, Andrea Facchinetti, and Giovanni Sparacino

Abstract— **Continuous glucose monitoring (CGM) sensors are minimally-invasive sensors used in diabetes therapy to monitor interstitial glucose concentration. The measurements are collected almost continuously (e.g. every 5 min) and permit the detection of dangerous hypo/hyperglycemic episodes. Modeling the various error components affecting CGM sensors is very important (e.g., to generate realistic scenarios for developing and testing CGM-based applications in type 1 diabetes simulators). In this work we focus on data gaps, which are portions of missing data due to a disconnection or a temporary sensor error. A dataset of 167 adults monitored with the Dexcom (San Diego, CA) G6 sensor is considered. After the evaluation of some statistics (the number of gaps for each sensor, the gap distribution over the monitoring days and the data gap durations), we develop a two-state Markov model to describe such statistics about data gap occurrence. Statistics about data gaps are compared between real data and simulated data generated by the model with a Monte Carlo simulation. Results show that the model describes quite accurately the occurrence and the duration of data gaps observed in real data.**

I. INTRODUCTION

Type 1 diabetes is a chronic, metabolic disease in which the pancreas produces little or no insulin; it is characterized by elevated levels of blood glucose (BG), which lead, if not treated, to serious damage to the heart, blood vessels, eyes, kidneys and nerves [1]. Treatment of diabetes involves diet, physical activity and accurate exogenous insulin administrations to keep the glucose level in safe range [2]. Since the 70s, the monitoring of BG at home has become possible thanks to self-monitoring blood glucose (SMBG) devices that measure the glucose concentration in a small drop of capillary blood collected by fingerprick. Since these measurements are collected about 3-4 times a day only, they are not able to detect all critical episodes of hypo/hyperglycemia occurring in daily life.

More recently, continuous glucose monitoring (CGM) devices were introduced that can measure almost continuously (e.g. every 5 min) glucose concentration in the interstitial fluid for several days/weeks [3][4]. The most popular CGM sensors are minimally-invasive electrochemical sensors that consist of a needle electrode placed in the subcutaneous tissue of the abdomen or the arm that measures a current signal originated by glucose oxidation, which is then converted to a glucose concentration profile using a conversion function. The glucose concentration readings of the sensor are finally transmitted to a receiver that displays the measurements in real-time to the patient [5]. As in any measurement device, the glucose values provided by CGM sensors are affected by errors. While a lot of work has been done to describe and model the major error components of the CGM sensor, such as the distortion introduced by the blood-to-interstitium kinetics, calibration, and random noise errors [4], existing mathematical descriptions of the occasional transient faults that affect CGM sensors need further investigation. In particular, one of the most common faults occurring in CGM data are data gaps, which are missing data due to an interruption of communication between the sensor transmitter and the receiver or to a temporary sensor error. A mathematical model of data gaps would be important both to characterize the occurrence of these faults and to mimic the generation of gaps in diabetes simulators [6].

The aim of this work is to develop a model of data gaps for a new-generation sensor, the Dexcom G6, developing further a pioneeristic approach proposed in Facchinetti et al. [7] for a sensor belonging to a previous technology.

M. Drecogna, M. Vettoretti, S. Del Favero, A. Facchinetti and G. Sparacino are with the Department of Information Engineering, University of Padova, Padova, 35131 Italy (phone: +39 0498277595; fax: +39 0498277699; email: {drecognama, vettore1, sdelfave, facchine, gianni}@dei.unipd.it).

II. DATASET

A. Dataset composition

The dataset was collected in 167 adults with diabetes, wearing the Dexcom G6 sensor for 10 days, which provides glucose concentration readings every 5 min. The data are part of the ones collected in adults during the Dexcom G6 Pivotal trial [8]. For this work, the raw CGM data collected in the Dexcom G6 Pivotal trial have been processed with an enhanced algorithm included in the newest version of the Dexcom G6 sensor, recently approved by the U.S. Food and Drug Administration, which enhances data availability. Since 36 patients wore two sensors in parallel, 203 CGM traces are available. For this analysis, only tracks with a minimum duration of 9 days are considered (172), in order to have homogeneous data.

B. Analysis of the data gaps

Data gaps are portions of missing data due to a disconnection or a temporary sensor error; an example of CGM trace with data gap due to temporary error is reported in Fig. 1.

Considering the available dataset, composed by 172 traces, we identified 229 gaps due to temporary sensor errors. Disconnections were not present in the considered dataset. A key point for the construction of the model of gaps concerns the analysis of their statistical distribution. We focused on the three gap statistics shown in Fig. 3: the number of gaps for each sensor, the distribution of gaps over the monitoring days and the data gap duration distribution. In the left panel, we can observe that about the 65% of traces have no gaps,

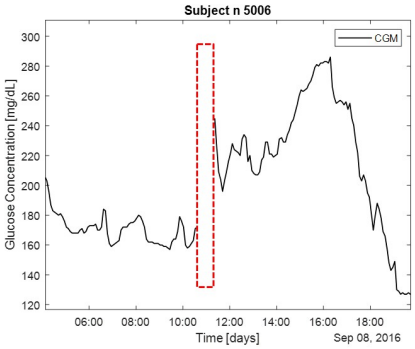

Fig. 1: Example of a portion of CGM data containing a gap between time 10:00 and 12:00 (marked with a red rectangle).

while most of tracks with gaps present only one event. In the middle panel it is possible to evidence that the frequency of data gaps varies with the time from sensor insertion, and in particular it is higher at the end of the sensor life. In the right panel the distribution of gap duration is reported: the median (5^{th}–95^{th} percentile) gap duration is 20 min (5 – 55 min).

III. MODEL OF SENSOR GAPS

A. Two-State Markov Model

The model proposed to describe the gap occurrence is a two-state Markov model which has been used in the past for describing gaps of the Dexcom G4 Platinum sensor [7].

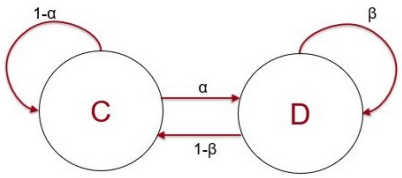

Fig. 2: A two-state Markov Model. In state C the system is regularly functioning, whereas in state D the sensor measurement is missing due to a data gap.

As shown in Fig. 2, the model is characterized by two states: the state C (where C stands for "connected") describes the normal functioning of the sensor, while the state D (where D stands for "disconnected") describes the presence of a data gap. Transitions between states are regulated by 4 probability values: if the sensor is normally functioning, α is the probability that the next measurement will be missing (C to D transition), while (1-α) the probability that the normal functioning will continue (C to C transition). Moreover, if the sensor is in state D, i.e. a data gap started, β is the probability that the data gap will continue for the next sample whereas the return to the sensor normal functioning, i.e. the end of the data gap, has probability (1-β). Calling d the duration of a data gap, the probability that a gap lasts for k samples, according to this model, is:

$$P([d=k]) = \beta^{k-1}(1-\beta) \tag{1}$$

that corresponds to the $k-1$ consecutive D-D transitions followed by a transition from D to C. The transition probabilities α and β can be estimated by maximum likelihood:

$$\hat{\alpha} = \frac{\text{\# of data gaps}}{\text{\# of regular samples}} \tag{2}$$

$$\hat{\beta} = \frac{\text{\# of missed samples preceded by a missed sample}}{\text{\# of missed samples}} \tag{3}$$

This first considered model (Model 1) is very simple; it is based on the hypothesis that both α and β are constant in time: this means that here the probability of having a gap and the probabilities of their duration do not change with time. Since these assumptions do not reflect the actual pattern of gaps (Fig. 3), a first adjustment that we can introduce is making α time-dependent, to improve the description of the distribution of gap occurrence in the various days of monitoring. To do that α is defined as a staircase function of time from sensor insertion t:

$$\begin{aligned} \hat{\alpha}(t) &= \alpha_k \ \text{ in day k from sensor insertion} \\ \hat{\alpha}_k &= \frac{\text{\# of data gaps in day k}}{\text{\# of regular samples in day k}} \end{aligned} \tag{4}$$

We can decide to use a different α value for each day of monitoring k (obtaining 10 different parameters) or to consider an α value for each group of days with similar gap probability. So far we adopted this second approach and we decided to group days 1,7 and 8 together as well as days 2 to 6, resulting in 4 different α parameters: $\alpha_{1,7,8}$, α_{2-6}, α_9, α_{10} (Model 2). Of course, this is an operator-dependent decision, modifiable according to preferences and needs.

B. Monte Carlo Simulation

The model is evaluated with a Monte Carlo approach, based on two steps:

1) The building of N = 100 simulated datasets, of the same size of the real one, in which data gaps are simulated with the identified model;
2) The comparison between the mean±SD of the data gap statistics obtained for the N=100 simulated datasets with the data gap statistics of the real dataset.

The gap simulation by the identified model (step 1) is shown in Fig. 4: for each simulated trace a sample x is extracted from a uniform distribution [0 1] and it is compared with the probability α of having a gap. If x is greater than α, there is no gap now and we have to pass to the next sample, extract another value x and again compare it with α.

This mini-cycle continues until the condition $x \leq \alpha$ is satisfied: in this case, the gap begins and so far it lasts for one sample only. To determine if the simulated gap will go on or not, another sample x_1 is extracted from a uniform distribution [0 1] and compared with the probability β that the gap continues: if the condition $x_1 \leq \beta$ is satisfied, the simulated gap continues, we pass to the next sample and we extract another value for x_1. This gap simulation advances

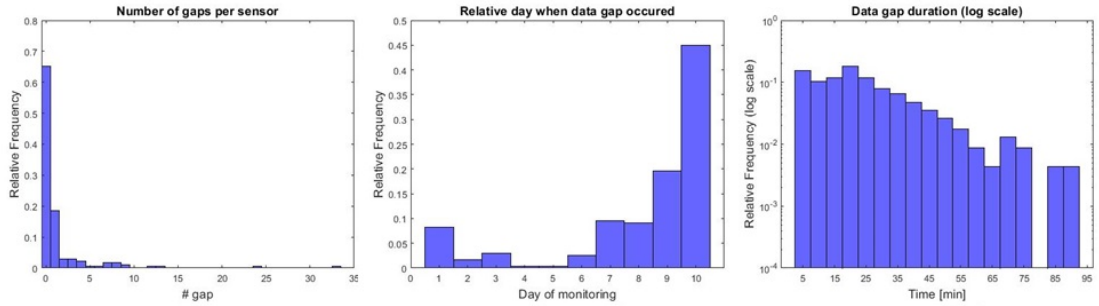

Fig. 3: Data gap statistics: Number of gap for each sensor (**left**), gap distribution over the monitoring day (**middle**), data gap duration (**right**).

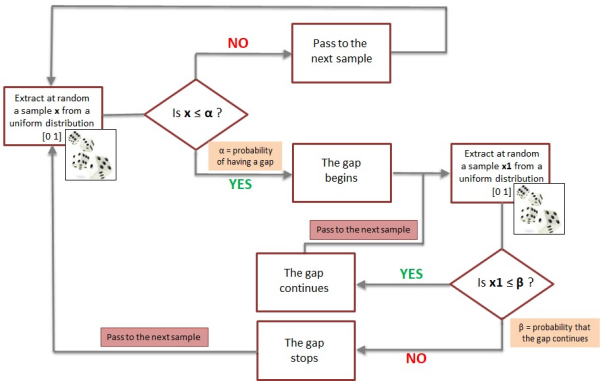

Fig. 4: Monte Carlo Simulation: if $x \leq \alpha$, the simulated gap begins. To know how many samples it will last, x_1 is compared with the probability β that the gap continues.

until x_1 becomes greater than β: in this case, the gap stops and with the next sample the cycle restarts. We presented the simulation for the simplest two-state Markov model with constant parameters, but the approach is the same for more complex Markov models with time-dependent parameters: the key point is to adapt α and β according to the specific case. For example, in the Model 2, α changes with the day since sensor insertion, so if we are simulating the first day of monitoring, we will refer to the $\alpha_{1,7,8}$ value, if we are simulating the 4^{th} we will take α_{1-6} and so on.

C. Results

The model performances are evaluated by comparing the mean±SD of the data gap statistics obtained for the simulated datasets with the data gap statistics of the real dataset. Fig. 5 reports the performances of the Model 1, whose parameters are constant over time and have been estimated from the data according to Eqs. (2), (3) (Table I). The red curves correspond to the mean result of the simulation, the whiskers represent the ± SD of the estimate and the blue histograms are the gap statistics in the real dataset.

In the left panel there is the description of the number of gaps for each sensor; we can observe that the fit is acceptable, yet not optimal. Indeed, the model estimates that almost the 30% of traces has no gaps, while in the real dataset about the 65% of CGM traces does not contain gaps. Moreover, the number of traces with 1, 2, or 3 events are overestimated by the model.

TABLE I: Model 1: α and β estimated values.

$\hat{\alpha}$	$\hat{\beta}$
4.65e-04	0.7082

In the middle panel the distribution of gaps over the monitoring days is represented. While in the real distribution the probability of having a gap in the last days from the sensor insertion is higher than in the first ones, the model simulates gaps uniformly over the different days. This is due to the fact that α is constant over time.

In the right panel the distribution of gap duration is reported; the log-scale is used in order to be able to appreciate the model performance also for low probability values, that corresponds to long gap durations. The model, being characterized by a single β parameter, fails to describe the peak of the distribution at 20 min, and simulates only a decreasing linear trend. In addition to the fit of the model, the panel shows the curve obtained from the theoretical formula that calculates the probability of having a gap that lasts for k samples depending on β (Eq. (1)): since this theoretical curve (green curve) and the result of the simulation (red curve) are overlapping, we can confirm that the simulation algorithm is well defined and also that 100 repetitions in the Monte Carlo simulation are sufficient to obtain the theoretical results.

As expected, the performances of Model 1 are not so good; therefore, we consider the second model, Model 2, that introduces a time-dependence for α. According to the eq. (4), the 4 parameters defining the α staircase function have been estimated and are reported in Table II. The β value for Model 2 is the same of Model 1 ($\hat{\beta}$ = 0.7082). This adjustment improves the description of the distribution of gaps over the monitoring days (middle panel in Fig. 6), whereas the description of the number of gaps per sensor and the data gap duration is comparable between Model 1 and 2.

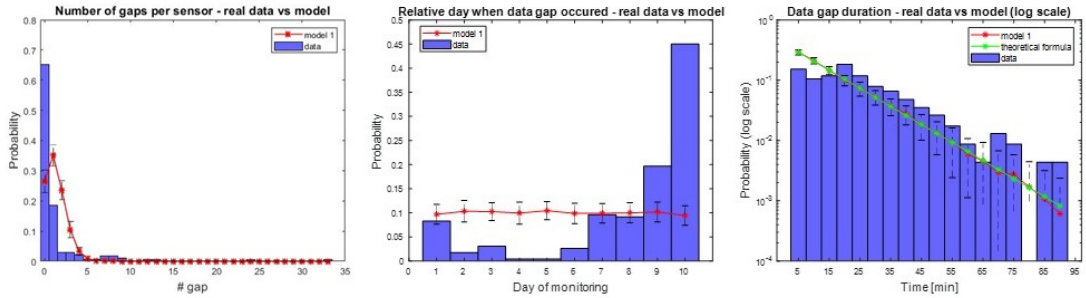

Fig. 5: Model 1 performances: Number of gaps for each sensor (**left**), gap distribution over the monitoring day (**middle**), data gap duration (**right**).The red curves correspond to the mean result of the simulation, the whiskers represent the interval mean $\pm$ SD of the estimate and the blue histograms are the gap statistics in the real dataset.

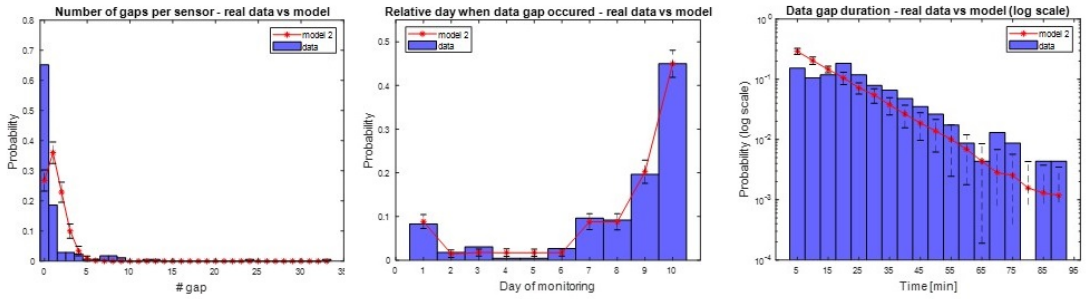

Fig. 6: Model 2 performances: Number of gaps for each sensor (**left**), gap distribution over the monitoring day (**middle**), data gap duration (**right**).The red curves correspond to the mean result of the simulation, the whiskers represent the interval mean $\pm$ SD of the estimate and the blue histograms are the gap statistics in the real dataset.

TABLE II: Model 2: α values estimated for each group of days.

$\hat{\alpha}_{1,7,8}$	$\hat{\alpha}_{2-6}$	$\hat{\alpha}_{9}$	$\hat{\alpha}_{10}$
4.06e-04	7.67e-05	9.11e-04	2.20e-03

IV. CONCLUSIONS

We developed a model for the description of data gaps caused by temporary sensor errors for the Dexcom G6 sensor in the adult population. We started from a simple two-state Markov model already used for the description of data gaps in past generation sensors, but its performances resulted not satisfactory for the Dexcom G6 sensor. Therefore, we improved the model by making α dependent on time. This second model was able to well describe the distribution of gap occurrence over the monitoring days. In future works, the addition of other states to the model will be investigated, in order to improve also the description of the distribution of gap duration and the number of gaps for each trace. Moreover, the model will be extended to the pediatric population that might present data gaps with different characteristics.

ACKNOWLEDGMENT

Dexcom Inc. (San Diego, CA) is acknowledged for having provided the dataset used in this paper. Dexcom, Dexcom G4, and Dexcom G6 are registered trademarks of Dexcom, Inc. in the United States and/or other countries.

REFERENCES

[1] World Health Organization (WHO). Overwiew, Sintoms and Treatment of Diabetes. Available online: https://www.who.int/health-topics/diabetes

[2] World Health Organization (WHO). Diabetes Facts Sheet. Available online: https://www.who.int/en/news-room/fact-sheets/detail/diabetes

[3] G. Cappon, G. Acciaroli, M. Vettoretti, A. Facchinetti, and G. Sparacino, "Wearable continuous glucose monitoring sensors: A revolution in diabetes treatment", *Electronics (Switzerland)*, vol. 6, no. 3, pp. 1–16, Sep 2017.

[4] M. Vettoretti, S. Del Favero, G. Sparacino, and A. Facchinetti, "Modeling the error of factory-calibrated continuous glucose monitoring sensors: Application to Dexcom G6 sensor data", *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2019, pp. 750–753, Jul 2019.

[5] M. Vettoretti, C. Battocchio, G. Sparacino, and A. Facchinetti, "Development of an error model for a factory-calibrated continuous glucose monitoring sensor with 10-day lifetime", *Sensors (Switzerland)*, vol. 19, no. 23, pp. 1–17, Dec 2019.

[6] M. Vettoretti, A. Facchinetti, G. Sparacino and C. Cobelli, "Type-1 Diabetes Patient Decision Simulator for In Silico Testing Safety and Effectiveness of Insulin Treatments," *IEEE Trans Biomed Eng*, vol. 65, no. 6, pp. 1281-1290, June 2018.

[7] A. Facchinetti, S. Del Favero, G. Sparacino, and C. Cobelli, "Modeling transient disconnections and compression artifacts of continuous glucose sensors", *Diabetes Technol Ther*, vol. 18, no. 4, pp. 264–272, Apr 2016.

[8] R.P. Wadwa, L.M. Laffel, V.N. Shah, S.K. Garg, "Accuracy of a factory-calibrated, real-time continuous glucose monitoring system during 10 days of use in youth and adults with diabetes", *Diabetes Technol Ther*, vol.20, no. 6, pp. 395-402, Jun 2018.